



(19)

Europäisches Patentamt
European Patent Office
Office européen des brevets



(11)

EP 0 955 749 A1

(12)

EUROPEAN PATENT APPLICATION

(43) Date of publication:
10.11.1999 Bulletin 1999/45

(51) Int Cl. 6: H04L 12/56, H04L 12/18

(21) Application number: 99303353.9

(22) Date of filing: 29.04.1999

(84) Designated Contracting States:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU
MC NL PT SE
Designated Extension States:
AL LT LV MK RO SI

(30) Priority: 08.05.1998 CA 2237208
08.05.1998 CA 2237264

(71) Applicant: NORTEL NETWORKS CORPORATION
Montreal, Quebec H2Y 3Y4 (CA)

(72) Inventors:
• Thoo, Joseph Fook-Kong
Saratoga, CA 95070 (US)

- Hadil Salim, Jamal
Nepean, Ontario K2B 7T2 (CA)
- Nandy, Blewajit
Ottawa, Ontario K2B 8J5 (CA)
- Seddigh, Nabil
Ottawa, Ontario K1S 5P3 (CA)

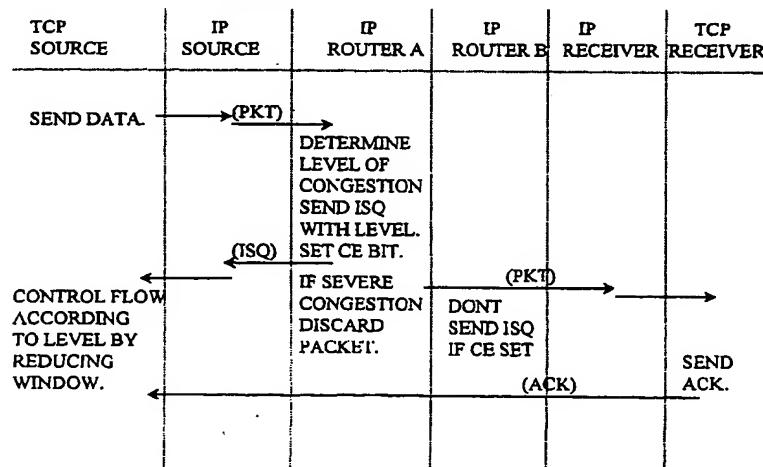
(74) Representative: Cutforth, Peter Nicholas
Nortel Networks
Intellectual Property Law Group
London Road
Harlow, Essex CM17 9NA (GB)

(54) Receiver based congestion control and congestion notification from router

(57) In a router (30) in a network comprising a source node (10), and a receiver node (20), and other nodes, a congestion monitor determines a degree of congestion, which is sent back to the source node, using an OSI network layer protocol. This enables the flow of packets from the source to be controlled more accurately to maintain high throughput with reduced probability of congestion. Using the network layer rather than lower

layers can ensure the indication can be carried across the entire network. The receiving host determines if the packet has been marked by any of the nodes on its path, to indicate congestion, e.g. by checking the CE bit in an IP header. A packet flow control parameter is generated at the receiving side, and sent to the source with the packet acknowledgment. This can reduce control loop delays compared to calculating congestion level by counting acknowledgments received at the source.

FIG. 4



EP 0 955 749 A1

Description**Background to the Invention****Field of the Invention**

[0001] The invention relates to packet routing apparatus for routing packets in a network, to source apparatus for use in a packet network, to methods of processing packets in a network, to receiver apparatus for use in a packet networks, to source node apparatus for use in Internet Protocol networks, to methods of processing packets in Internet Protocol networks, and to corresponding software.

Background art

[0002] Congestion in packet networks may occur for example at routers where flows converge from different sources. As complicated networks may consist of numerous different end systems (sources and receivers), routers, and links, it is usually impossible to match their capacities perfectly. Accordingly, congestion will occur where more packets are received than can be handled. Various ways of handling congestion are known. At the simplest level, buffering is provided to handle temporary overloads. For longer overloads, flow control mechanisms are provided, to enable downstream elements to cause the source to reduce the rate of sending packets.

[0003] If a buffer overflows before the source reduces the flow, then packets will be discarded. The source may eventually detect that these packets have not been acknowledged, and retransmit them. This can make the congestion worse, and lead to complete collapse of the network. On the other hand, if the flow control is made too strong to avoid such a collapse, the throughput of the network may be very low, thus making inefficient use of resources.

[0004] Other mechanisms to improve efficiency include admission control, preallocation of buffers, and delay sensitive routing, to avoid congested regions.

[0005] Flow control relies on some sort of notification of congestion. This may be implicit or explicit. Congestion can be inferred for example from detecting at the source that packets have been discarded downstream, or detecting delays in the time taken for a packet to arrive. However, with such methods there may be a considerable lag between the congestion occurring, and it being detected. Also, it is possible that the packets were discarded or delayed for reasons other than congestion, e.g. link failure, or erroneous routing. Accordingly, explicit congestion notification mechanisms have been used. One method is described in US patent 5,377,927 (Jain et al) in the context of a system in which at intermediate nodes, a flow is allocated a share of the capacity. If the allocation is exceeded, a flag is set in each packet. The flags may be counted and if the proportion of packets with the flag set exceeds a threshold, then

the flow from the source is adjusted.

[0006] Another example is in Frame Relay, a data link layer protocol which has both forward explicit congestion notification (FECN) for notifying the receiver, and backward explicit congestion notification for notifying the source directly. ATM also has an FECN mechanism. The Internet Protocols (IP) also include an FECN and a BECN mechanism. The BECN mechanism is in the form of an Internet Control Message Protocol (ICMP) message called the ICMP Source Quench (ISQ) message. However, it is currently recommended that this message not be used, as it may consume too much bandwidth, and thus contribute to the congestion, and is unfair and ineffective in determining which of multiple flows should be limited.

[0007] It has been proposed that ISQs be used in conjunction with random early detection (RED) routers, to enable the rate of sending ISQs to be limited, and reflect how much each flow is contributing to the congestion. However, this has not been adopted, and instead, current practice in TCP/IP (Transport Control Protocol/Internet Protocol) involves using TCP, a transport layer protocol, to determine either lost packets or increases in delays using a timeout mechanism, or determining increases in delays, by timing the acknowledgment sent back by the TCP receiver. This enables the TCP sender to infer congestion and react by reducing its window for that flow, that is, the number of packets it can send towards a given receiver before it must wait for acknowledgments from the receiver.

[0008] Floyd [Sig94 paper "TCP and Explicit Congestion Notification"] discloses a methodology for doing ECN for IP [and later in an IETF draft, Nov 1997]. Floyd suggests the use of RED gateways to detect incipient congestion before the queue overflows. The congestion causing packets are marked on their way to the receiver end system (from the sender end system), with a probability proportional to their bandwidth usage, using the CE (Congestion Experienced) bit in the IP header. When the receiver end system receives the congestion causing packet they inform the sender end system to slow down when ACKing that packet by setting the ECE (Explicit Congestion Echo) bit in the TCP header.

[0009] The use of ECN capable gateways for notifying end systems prevents unnecessary packet drops. In the ideal situation where everyone supports ECN at the end system as well as at the intermediate nodes, sources are now informed quickly and unambiguously that there is congestion in the network and therefore do not have to rely on extrapolating that condition.. Floyd also suggests that with the use of tagging congestion-causing packets, other types of networks that the IP packet traverses e.g. ATM can employ their own congestion control algorithms as well as have the ability to mark congestion causing packets.

Summary of the Invention

[0010] It is an object of the invention to provide improved methods and apparatus.

[0011] According to a first aspect of the invention there is provided a packet routing apparatus for routing packets in a network comprising a source node, and a receiver node, and other nodes, the routing apparatus comprising:

input means for receiving a packet passed across the network from the source node;
a congestion monitor for determining a degree of congestion at the routing apparatus; and
output means coupled to the input means and to the congestion monitor, for sending an indication of this degree of congestion to the source node, using an OSI network layer protocol.

[0012] An advantage of sending an indication of the degree of congestion is that the flow of packets from the source can be controlled more accurately to maintain high throughput with reduced probability of congestion. An advantage of sending it straight back to the source rather than using the packet to carry the information on to the receiver then returned to the source, is the speed of response. The combination of speed and graded congestion information work together to enable the probability of severe congestion to be effectively reduced. An advantage of sending the indication at the network layer rather than higher layers is that the mechanism is not tied to any particular higher layer protocol. An advantage of sending at the network layer rather than lower layers is that it can ensure the indication can be carried across the entire network, and not be lost at boundaries between data links making up the network. Furthermore, as the receiver need not be involved in the congestion notification, there is no need for it to be ECN capable, and thus there is no need for a negotiation of ECN capability at flow set up time.

[0013] Preferably, the network layer protocol is an Internet Protocol.

[0014] Preferably the Internet Protocol used is the Internet Control Message Protocol Source Quench message.

[0015] Preferably the indication to the source is made proportional to how much the packets from the source contribute to the congestion, relative to packets from other nodes.

[0016] An advantage of this is that flow control fairness and effectiveness in preempting congestion can be improved, if the nodes sending most packets can be controlled to reduce their flow most, or first.

[0017] Preferably the routing apparatus further comprises an output rate adapter for making the indication proportional by adapting the rate of sending indications.

[0018] Preferably the routing apparatus further comprises a packet queue, the congestion monitor being ar-

ranged to operate according to how full is the packet queue.

[0019] An advantage of this is that it is easy to measure, and can enable queue overflow to be predicted and prevented.

[0020] Preferably the routing apparatus further comprises a packet marker means for marking the packet to indicate it has experienced congestion.

[0021] An advantage of this is that it can enable the receiver learn of congestion, and thus perhaps contribute towards solving it, e.g. by assisting in flow control. A further advantage of this is that it can improve compatibility with other nodes using different congestion notification methods. Furthermore, it enables subsequent nodes to suppress sending their congestion indications back to the source, if they know one has already been sent for that packet. This can reduce the bandwidth used by sending such indications, which would otherwise contribute to the congestion.

[0022] Preferably the routing apparatus further comprises means for determining from the packet, if it has previously triggered a sending, to the source node, of an indication of congestion, the output means being operable according to whether such an indication had been sent previously.

[0023] This can reduce the bandwidth used by sending such indications, which would otherwise contribute to the congestion.

[0024] Another aspect of the invention provides a source apparatus as set out in claim 9.

[0025] Another aspect of the invention provides a method of processing packets as set out in claim 10.

[0026] Preferably the method further comprises the steps of claim 11.

[0027] Another aspect of the invention provides a method of processing packets as set out in claim 12.

[0028] Preferably the method further comprises the step of receiving from the receiver node a flow control message, the step of controlling the flow of further packets being made also on the basis of this flow control message.

[0029] Another aspect of the invention provides packet routing apparatus as set out in claim 14.

[0030] An advantage of making sending the indication to the source node dependent on whether one has been sent previously, is that any addition to the congestion by sending the indications, can be reduced. This is particularly desirable in instances where there is congestion at multiple nodes, in which case, sending multiple indications may be avoided.

[0031] Another aspect of the invention provides a method of processing packets as set out in claim 15.

[0032] Preferably the packet is an Internet Protocol packet, and the step of determining from the packet, if it has previously triggered a sending of an indication comprising checking the Congestion Experienced bit in the packet header.

[0033] Preferably, the step of sending an indication is

not carried out if the indication had been sent previously, unless the routing apparatus discards the packet.

[0034] Another aspect of the invention provides software stored on a computer readable medium for carrying out the above methods.

[0035] According to another aspect of the invention there is provided a receiver apparatus for use in an Internet Protocol network comprising a plurality of nodes, the apparatus comprising:

input means for receiving a packet sent across the network using the Internet Protocol; packet reading means for determining if the packet has been marked according to the Internet Protocol by any of the nodes through which it passed, to indicate congestion at that node; a packet flow control parameter generator responsive to the packet reading means, for determining a packet flow control parameter; and output means for sending a message to the source, to control the flow of packets from the source, according to the packet flow control parameter.

[0036] An advantage of enabling the receiver to contribute to the flow control is that it can reduce control loop delays caused by waiting at the source for a number of acknowledgments to arrive. Also, it can improve reliability since it enables the receiver to force a rogue source to slow its flow of packets. A further advantage is that the receiver may be aware of different local information and/or conditions to those of the source, or may have a more up to date control algorithm, so overall effectiveness of the flow control may be improved if it is adapted by the receiver.

[0037] Preferably, the sending means is arranged to send the packet flow control parameter in the message.

[0038] An advantage of this is that the receiver may have more direct control over the flow.

[0039] Preferably, the message is an acknowledgement of receipt of the packet.

[0040] An advantage of this is that it helps keep the number of packets returned to the source, to a minimum. Furthermore, since many protocols already cater for sending acknowledgements, in many cases it will be easy to adapt them in this way, and there may be no additional bandwidth taken up by the message.

[0041] Preferably, the message is an acknowledgement of receipt of the packet, and the output means is arranged to delay the acknowledgement on the basis of the packet flow control parameter.

[0042] An advantage of this is that the receiver can enter the control loop with little or no modification to the source if it is already arranged to carry out flow control on the basis of acknowledgements received in a given time period. Hence this may give better backward compatibility.

[0043] Preferably the packet flow control parameter generator is arranged to be dependent additionally on

whether preceding packets were marked.

[0044] This enables the control to be more steady in response to transient congestion for example.

[0045] Preferably the packet flow control parameter comprises an 'offered' window size, for indicating to the source node how many packets can be sent before the source should wait for an acknowledgement from the receiver.

[0046] An advantage of using this parameter is that existing apparatus is arranged to negotiate window size as part of flow set up, and so can be adapted more easily to this new arrangement.

[0047] Preferably the message is an acknowledgement of receipt of the packet, the output means being arranged to send the same packet flow control parameter again in an acknowledgement sent in response to a subsequent packet.

[0048] An advantage of this is that it may reduce the risk of flow control being affected by the acknowledgements being delayed or lost in transit.

[0049] Preferably the packet flow control parameter generator is operable according to redefinable parameters.

[0050] An advantage of this is that the flow control can be updated or tailored more easily.

[0051] Preferably the packet flow control parameter generator is operable additionally according to a receiver-specific parameter.

[0052] This enables different receivers to control the flow to suit their local conditions better.

[0053] Another aspect of the invention provides a source-node apparatus for use in an Internet Protocol network comprising a plurality of nodes, the apparatus comprising:

35 a packet sending means for sending a packet using across the Internet Protocol network to one of the nodes acting as a receiving node; a receiver for receiving from the receiving node a packet flow control parameter sent across the Internet Protocol network; and 40 a flow controller coupled to the packet sending means for controlling a rate of flow of sending further packets from the sending means to the receiving node on the basis of the packet flow control parameter.

[0054] Preferably the receiver is capable of receiving congestion notification messages from other nodes in 50 the path of the packet between the source node and the receiving node, the flow controller being operable to control the rate of flow on the basis of the congestion notification messages received from the other nodes.

[0055] Another aspect of the invention provides a 55 method of processing packets in an Internet Protocol network comprising a receiver node, a source node and intermediate nodes, the method comprising the steps of: at the receiver node,

receiving a packet sent across the network using the Internet Protocol;
determining if the packet has been marked by any of the nodes through which it passed, to indicate congestion at that node;
determining a packet flow control parameter on the basis of the determination of a marking; and sending a message to the source node across the Internet Protocol network, to control the flow of packets from the source node, according to the packet flow control parameter.

5

10

[0056] Preferably the method further comprises the steps of: at the source node,

15

receiving the message; and controlling a rate of flow of sending further packets to the receiving node on the basis of the message.

20

[0057] Preferably the message contains the packet flow control parameter, the controlling step being made on the basis of the parameter.

[0058] Preferably the method, further comprises the step of delaying sending the message from the receiver node, on the basis of the flow control parameter, the flow controlling step being made by the source on the basis of the time of arrival of the message.

[0059] Preferably the method further comprises the preliminary step of determining if the intermediate nodes are able to send congestion notification messages to the source node, the step of determining the flow control parameter at the receiver being carried out on the basis of whether the intermediate nodes are capable of sending congestion notification messages to the source.

[0060] Preferably a method uses an application to transmit data across a network, the application causing the network to use a receiver apparatus as set out above to transmit the data.

[0061] Another aspect of the invention provides software on a computer readable medium for carrying out the above methods.

[0062] Any of the preferred features may be combined, and combined with any aspect of the invention, as would be apparent to a person skilled in the art. Other advantages will be apparent to a person skilled in the art, particularly in relation to prior art other than that mentioned above.

[0063] To show, by way of example, how to put the invention into practice, embodiments will now be described in more detail, with reference to the accompanying drawings.

Brief Description of Drawings

[0064]

FIGURES 1 to 3 show known arrangements;
FIGURE 4 shows in sequence chart form the ac-

tions of elements in a TCP/IP network corresponding to that shown in FIGURE 1, according to an embodiment of the invention showing sending a congestion level in an ISQ;

FIGURE 5A shows parts of a RED process in a router carrying out the actions of FIGURE 4;

FIGURE 5B shows parts of an ISQ send process in a router carrying out actions of FIGURE 4;

FIGURE 6 shows an example of flow control actions of FIGURE 4 of a TCP source receiving an ISQ;

FIGURE 7 shows in sequence chart form the actions of elements in a TCP/IP network corresponding to that shown in FIGURE 1, according to another embodiment of the invention showing determining window size at the receiver;

FIGURE 8 shows an example of flow control actions of FIGURE 7 of the TCP receiver;

FIGURE 9 shows an example of flow control actions of FIGURE 7 of the TCP source;

FIGURE 10 shows in sequence chart form the actions of elements in a TCP/IP network corresponding to that shown in FIGURE 1, according to another embodiment of the invention showing delaying ACK at the receiver; and

FIGURE 11 shows an example of flow control actions of FIGURE 4,7 or 10 of a TCP source receiving an ACK with the ECN notify bit set.

Detailed Description

30

FIGURES 1-3, prior art

[0065] FIGURE 1 shows in schematic form some of the principal elements in a network using TCP/IP. A host A,10 is being used to send data to another host B,20, across an IP network. An application running on host A, or remotely, delivers the data to be transmitted to a TCP source endpoint, 50. This passes the data to IP source functions 60. The source functions send IP data packets

to an IP router A,30. The packets are routed via another IP router B,40, and eventually reach the IP receiver functions 80 on host B. The IP receiver functions 80 demultiplex the IP packets and pass the data to the TCP receiver endpoint 70 in the host B. The TCP receiver

transfers the data to higher level software in the host B. [0066] As indicated by the vertical arrows to the IP routers A and B, 30 and 40, many other paths may converge at the routers. Queues 90 are provided, to handle transient congestion. In practice there may be many

more than three hops between source and receiver. TCP/IP is shown as an example, but there are many other protocols for which a similar arrangement would apply, both in OSI layers 4 and 3 respectively, and in other layers.

[0067] FIGURE 2 shows some of the principal actions of each of the elements of FIGURE 1, for the proposal mentioned above that ISQ messages be used as a congestion notification system. The TCP source sends data

via the IP source to router A. If there is severe congestion at router A, the packet may be discarded. If the router detects incipient congestion, an ISQ is sent back to the IP source. The IP source will interpret the ISQ and pass an indication up to the TCP source functions, notifying of the congestion at router A, on the flow to the TCP receiver at host B.

[0068] The TCP source functions react by reducing the window to control the flow. If the packet is not discarded by router A, it is forwarded on via router B to the IP receiver and ultimately to the TCP receiver. The TCP receiver sends an acknowledgement back to the TCP source.

[0069] FIGURE 3 shows an example of the above-mentioned proposal by Floyd. Again the actions of the principal elements of figure 1 are shown. There may be an initial negotiation before flow starts, between the TCP source and the TCP receiver. The source may query if the TCP receiver is ECN capable. If the TCP receiver replies positively, e.g., by using the ECN notify bit (also called the ECN-ECHO bit) in the SYN packet or the SYN-ACK packet.

[0070] If both TCP source and TCP receiver are ECN capable, the TCP source instructs the IP source to set the ECT bit in the IP header of data packets for that flow. If the router A detects incipient congestion, it may either discard the packet, if the congestion is too severe, or, if less severe, it can mark the packet by setting the CE bit in the IP header. When a marked packet is received by the IP receiver, it will notify the TCP receiver of the congestion notification. The TCP receiver will then send an ACK with the ECN notify bit set. This is sent back to the TCP source, which reacts by reducing the window size to control the flow. The source does not respond to further ECN ACKs until the end of that window.

[0071] If no ACK is received for any reason, e.g., router A has discarded the packet, after a time-out, the TCP source retransmits the packet.

[0072] If a packet is received at the source with ECN-notify set in the TCP header then the source knows that there is network congestion and reacts by halving both the congestion window, cwnd and the slow start threshold value, ssthresh.

[0073] The source does not react to ECN more than once per window. Upon receipt of an ECN-notify packet at time t, it notes the packets that are outstanding at that time (sent but not yet acked, snd_una) and waits until a time u when they have all been acknowledged before reacting to a new ECN message. The sender does not increase the congestion window in response to an ACK if the ECN-notify bit is set. Incoming acks will still be used to clock out data if the congestion window allows it. TCP still follows existing algorithms for sending data packets in response to incoming ACKs, multiple duplicate ACKs or retransmit timeouts.

FIGURE 4 - embodiment using ISQ

[0074] FIGURE 4 shows the actions of elements in a TCP/IP network corresponding to that shown in FIGURE 1. The TCP source sends a data to the IP source, which sends it in the form of IP packets to router A. Router A determines the degree of congestion. It discards the packet if congestion is very severe, and sends an ISQ back to the IP source. For other levels of congestion, it sends an ISQ to the source indicating the level of congestion, and marks the packet by setting the CE bit, before passing it on to router B. Router B does the same except that if the packet was marked by a preceding router, it does not send a further ISQ, since the TCP source has already been alerted by the first ISQ. An exception is where the congestion is more severe in IP router B. For example, if the packet is discarded at IP router B, then sending an ISQ from router B to TCP source may be justified, to enable more drastic flow control to be implemented if desired.

[0075] The IP receiver, unlike the prior art case shown in figure 3, does not need to be ECN capable, since the ISQ notification is enough for flow control. Nevertheless, if the receiver is ECN capable, perhaps for compatibility, it would detect whether the CE bit and ECT bit are set in the IP header, and notify the TCP receiver of the congestion indicated by these bits. As in FIGURE 3, the TCP receiver sends an ACK with the ECN notify bit set, back to the TCP source. This may or may not be used by the source, in addition to the notification from the ISQs. As will be discussed in more detail below, the TCP source will control the flow of packets by reducing its window, according to the level of congestion indicated in the ISQ. If it can also control the flow according to whether ACK packets received have their ECN notify bit set or not, it may be advantageous, for compatibility with intermediate router nodes which cannot send ISQs.

FIGURES 5A,5B - example of router actions of FIGURE 4

[0076] FIGURES 5A and 5B show in more detail two processes happening inside the router in the embodiment of FIGURE 4: Firstly a random early detection (RED) process, and secondly an ISQ sending process, which may be invoked by the RED process. The RED process is an example of a congestion monitor. It is known and well documented, and need not be described here in detail. It monitors average queue lengths using a low pass filter. Many other methods could be used. The ISQ is an example of a mechanism for sending an indication of this degree of congestion to the source node, using an OSI network layer protocol. Other mechanisms can be used for this and other protocols.

[0077] In the RED process, at 200, an incoming packet arrives. At 210 the process determines if the average queue length is greater than a predetermined maximum threshold. If so, the packet may be discarded and the

ISQ sending process invoked.

[0078] Otherwise, at 230, it is determined if the average queue length is greater than a minimum threshold. If so, again, the ISQ sending process may be invoked, at 240, depending on whether the packet is chosen by the RED process for marking, 250. Only a proportion of packets are marked at this stage. The proportion is made dependent on the relative bandwidth used by each flow. This is designed to improve the fairness of the congestion notification, and make the overall flow control for many paths, more stable. This is one example of a mechanism for making the indication to the source proportional to how much the packets from the source contribute to the congestion, relative to packets from other nodes. Other examples can be used to achieve the same advantage, either based on sending messages selectively, i.e. limiting the rate of sending indications, or by other methods, such as indicating explicitly in the indication, and/or the marking, what is the proportion of the congestion attributable to a given flow, or connection.

[0079] At 260, the packet is placed in an appropriate queue in the router. This also takes place if the average queue size is less than the minimum threshold.

[0080] The ISQ send process begins at 270 by checking if the CE bit has been set previously. If not, at 280, a congestion level index is derived from the congestion level in the router, obtained by the RED process from the size of the queues. At 290 the packet source address is obtained, and the ISQ generated using the packet source address and the congestion level index. There is an unused 32 bit field in the ISQ which can be used for the level index. It is sent back to the IP source, and the process returns at 310.

[0081] If the CE bit has been set previously, at 320, it is determined if the packet has been dropped by the RED process. If not, the process returns. Otherwise, the congestion level index is set at a maximum, eight in this case, and the ISQ is generated and sent on the basis of this index and the packet source address. This means where there are multiple congested routers in the path, they will be dealt with in order, starting with the one closest to the source. Any bias in this will be mitigated by having an ISQ sent if a second or subsequent congested router discards a packet. This will contribute to the flow control at the source. If more balance is warranted, downstream routers could be allowed to send ISQs in some circumstances when they are not discarding a packet, but they find the CE bit already set.

FIGURE 6 - TCP source response to ISQ

[0082] FIGURE 6 shows in more detail the actions of a TCP source in the embodiment of FIGURE 4. The TCP source may be arranged to respond to congestion notification from either the intermediate nodes, or from the receiver, or both. In the case of being responsive to both, it could be arranged to respond to either type in-

dividually, or be arranged to control a flow on the basis of both types of notification simultaneously. The example of response to ISQs will now be described in more detail.

- 5 [0083] At 510, an ISQ is received for this flow. Different flows, for each different source-receiver pair, will have their own process. If it is determined at 520 from the ISQ that the congestion level is severe, e.g. level 8 in this example, then a rapid response is made. Otherwise, a more measured, gradual control algorithm is carried out.
- 10 [0084] The rapid response, 480, involves reducing the window by half and reducing the value of STHRESH by half. This value represents a threshold for the TCP slow starting procedure. This defines how the window is increased in size following a slow start, which involves reducing the window to one. At 490, the value of ECN_ACC is reset following this drastic window reduction. If it was caused by a level 8 ISQ, the source infers that the packet was dropped, and retransmitted at 500.
- 15 [0085] The example of the more gradual response involves incrementing a counter ECN_ACC, by the level of the congestion, 530. The counter reflects the number and severity of the congestion notifications received over a time period for a given flow. The flow is controlled by adjusting the window size. The window size indicates how many packets can be sent before the source must wait for an acknowledgment from the receiver. The source is constrained not to change the window size before the end of the current window, to meet standards, and ensure compatibility with receivers. Hence the counter may be incremented by numerous ISQs from successive packets, before the end of the window. The process returns at 450 to await the end of the window.
- 20 [0086] The counter need only be compared, 460, to a threshold, at the end of the window. If the value of ECN_ACC is below the threshold, at 470 the window is adjusted less drastically, depending on the value of ECN_ACC. Exactly how the window is adjusted is a matter for particular implementations. It could be incremented by one if the value of ECN_ACC is less than four (<4) but greater than zero (>0). If less than zero (<0), the window could be exponentially incremented. If between four and eight, the congestion window might be left unaltered. If greater than or equal to eight (> or =8) the more drastic congestion reaction at 480 mentioned above, to cut the window and the value of STHRESH by half would be carried out.
- 25 [0087] In another example, the source could react as described above in relation to FIGURE 3, and react immediately to an incoming notification, without accumulating a score. This could be appropriate if the ISQ notification is implemented without the level of congestion indication, but with the selective sending of ISQs for a proportion of the packets, and with the preliminary check of the CE bit.

FIGURE 7 - embodiment using receiver based flow control

[0088] FIGURE 7 shows actions in a similar manner to the diagram of FIGURE 4. There are two significant distinctions. Firstly, the TCP receiver, instead of sending an ACK with the ECN bit set, instead determines an offered window size and sends this with the ACK. Secondly, at the TCP source, instead of determining a revised window size on the basis of ECN notify bits in the ACK packets, instead takes the offered window size from the ACK signal, and uses that. The hardware or software in the IP receiver which is arranged to check the CE bit in the IP packet header, is an example of a packet reading means for determining if the packet has been marked by any of the nodes through which it passed, to indicate congestion at that node. The hardware or software in the TCP receiver, or invoked by the TCP receiver, is an example of a packet flow control parameter generator responsive to the packet reading means, for determining a packet flow control parameter.

[0089] The offered window size is one example of a parameter for controlling the flow. Others can be conceived. Doing the window size calculation at the receiver has a number of benefits. Firstly, there will be a faster response, because the source need not wait for all ACKs, before being able to determine accurately whether the proportion which have the ECN notify bit set is sufficient to change the window size. Secondly, redundancy is built in because many ACKs will be sent with a calculated window size. Thirdly, the accuracy of the calculation of the window size is not dependent on successful transmission of all the ACKs.

[0090] The benefit of faster response can be illustrated by an example of a worst-case scenario. If a group of ACKs are delayed by varying amounts so that some arrive at the source near the end of a window and others arrive later, if the source makes the window calculation, as in the prior art, it may determine that the next window be unaltered or larger, as only a few ACKs with ECN notify bits have arrived in time. In contrast, if the receiver makes the calculation, as soon as enough packets have arrived at the receiver with their CE bit set, thereafter all ACKs issued by the receiver will contain the reduced offered window size. These will be subject to varying delays in their path back to the TCP source. However, the fastest of them will trigger the window reduction at the TCP source. If the fastest one arrives before the end of the previous window, then the next window will be reduced in size, and thereby the amount of congestion reduced. This shows how receiver-based window calculations may give a faster response than simply sending back ECN notify bits and allowing the TCP source to make the window size calculation.

FIGURE 8 - receiver actions in the example of FIGURE 7

[0091] FIGURE 8 shows in more detail the actions of the receiver in response to the arrival packets, according to the example of FIGURE 7. When a packet is received, at 600, the flow is identified, from the source address field. At 610, the data in the packet is processed. Simultaneously or afterwards, at 620, there is a test for whether the packet has experienced congestion, as indicated by the ECN and ECT bits in the IP packet. If so, at 630, the value of ECN.RCVD, is incremented. This indicates an accumulation of how many packets have experienced congestion. Otherwise, if the packet has not experienced congestion, at 640, the same value is decremented.

[0092] In either case, the value of ECN.RCVD is tested to see if it is above a given threshold at 650. If so, drastic reduction in window size is made. Otherwise, a gradual alteration in the window size can be made, according to the precise value of ECN.RCVD, at 670.

[0093] The new window size is then output as an offered window size in a field in the ACK packet. This is sent to the TCP source, at 680.

[0094] Just how the window size or other control parameter is calculated in the receiver need not be specified or limited by a standard. This would enable future enhancements to be carried out without having to alter standards, or raising compatibility problems. The calculation could be carried out as part of a separate software module which could be redefined independently of other modules.

[0095] This would make it easier to tailor the calculation to take account of receiver specific parameters. For example, the receiver might have a local policy if it is in a different network domain to the source. Local conditions such as local quality of service policies, or local network characteristics may dictate a different window size calculation algorithm. For example, part of the network may be implemented over wireless connections, or satellite connections, in which case different window sizes and different control of the window size may be appropriate to take account of the different delay characteristics.

FIGURE 9 - TCP source actions for the example of FIGURE 7

[0096] FIGURE 9 shows an example of flow control actions of FIGURE 7 of the TCP source, where the receiver contributes to the flow control. At 410, an ACK is received by the source, for the given flow. Again different flows would have their own processes, windows and counters as appropriate. At 800, it is determined if the ACK contains a new offered window size, in the appropriate field in the ACK packet. If not, at 840, the source may check for an ECN notify bit, and process it as described in more detail below with reference to FIGURE

11.

[0097] If there is an offered window from the receiver, this is stored for use at the end of the current window, at 810. At 820, the process returns to await more ACKs, until the end of the current window. Then at 830, the stored offered window is used for the new window. It would be possible to calculate a source window and take the smaller of the source calculated and receiver calculated window sizes, if it is desired to share the influence on the flow control between the source and the receiver. [0098] Where a number of offered windows of different sizes are sent back by the receiver, the source could choose the smallest, or some other selection algorithm could be implemented.

FIGURE 10, alternative embodiment using receiver based flow control

[0099] FIGURE 10 shows in sequence chart form the actions of elements in a TCP/IP network corresponding to that shown in FIGURE 1, according to another embodiment showing flow control by delaying ACK at the receiver. This shows how the IP receiver may check the CE bit as in the example shown in FIGURE 7, but the actions of the TCP receiver and TCP source differ from that example. Instead of adding a window size to the ACK, or setting the ECN notify bit, as in FIGURE 3, the TCP receiver determines a delay for the ACK, based on the congestion indications from the current packet, and from preceding packets. The ACK is sent after the delay, and the TCP source reacts accordingly.

[0100] The TCP source reacts to delayed ACKs by assuming that the delays are being caused by congestion, and so reduces its flow rate by reducing the window size. As it does not rely on the source processing any ECN bits, this has the advantage that it can be used with non ECN capable sources, and needs no ECN negotiation when setting up a flow. Furthermore, there is no reliance on particular fields in a TCP packet header, so in principle it can be used with other layer four protocols running over IP. The receiver could be arranged to set the ECN notify bit as well, for the source to use if it is ECN capable, since these may enable the receiver to control the flow more accurately, since they will make the source response less dependent on any transit delays for the ACKs.

FIGURE 11, - alternative TCP source actions for the example of FIGURE 4,7 or 10

[0101] Figure 11 shows an example of flow control actions of FIGURE 4,7 or 10 of a TCP source receiving an ACK with the ECN notify bit set. This response may be arranged to occur in conjunction with or as an alternative to flow control based on other inputs. If an ACK is received at 410, the TCP source determines at 420 if the EC notify bit has been set. If so, an accumulating count of congestion notifications, labelled ECN_ACC is incre-

mented at 430. Otherwise, at 440, if an ACK is received without the EC notify bit being set, the value of ECN_ACC is decremented at 440. The process returns at 450 to await the end of the window. At the end of the

5 window, the window adjustment may be made in the same way as described in relation to FIGURE 6, reference numerals 460 - 500. By accumulating a score of congestion notifications, instead of reacting to the first notification in each window, as described with respect to figure 3, better control can be achieved. For example 10 transient congestion once per window may not merit reducing the window size.

Hardware examples.

15 [0102] In principle, the source, receiver, router, and other elements described above, could be implemented on a wide range of different types of well known hardware. If speed of operation or capacity are not overriding factors, the OSI layer 3 and 4 functions could be implemented as software processes on a general purpose workstation. Dedicated hardware could be used for specific functions to improve speed, as would be well known. Protocol specific hardware such as linecards, 20 and physical transmission links between remote computers would be used for lower OSI layer functions. The choices are matters of implementation following established principles, and need not be described further here.

25 [0103] **Other Variations**

30 [0104] Although in the embodiments described, TCP is used, other similar OSI layer four protocols may be used. Likewise, although examples have been described using IP, other OSI layer three protocols may be used as appropriate. The intermediate nodes have been described using the example of a router, though this should not be read as excluding other intermediate 35 nodes where congestion could occur.

[0105] Other variations within the scope of the claims will be apparent to persons of average skill in the art, and are not intended to be excluded.

Claims

1. Packet routing apparatus for routing packets in a network comprising a source node, and a receiver node, and other nodes, the routing apparatus comprising:

55 input means for receiving a packet passed across the network from the source node; a congestion monitor for determining a degree of congestion at the routing apparatus; and output means coupled to the input means and to the congestion monitor, for sending an indi-

- cation of this degree of congestion to the source node, using an OSI network layer protocol.
2. The routing apparatus of claim 1 the network layer protocol being an Internet Protocol. 5
3. The routing apparatus of claim 2 the Internet Protocol used being the Internet Control Message Protocol Source Quench message. 10
4. The routing apparatus of claim any of claims 1 to 3 wherein the indication to the source is made proportional to how much the packets from the source contribute to the congestion, relative to packets from other nodes. 15
5. The routing apparatus of claim 4 further comprising an output rate adapter for making the indication proportional by adapting the rate of sending indications. 20
6. The routing apparatus of any of claims 1 to 5, further comprising a packet queue, the congestion monitor being arranged to operate according to how full is the packet queue. 25
7. The routing apparatus of any of claims 1 to 6, further comprising a packet marker means for marking the packet to indicate it has experienced congestion. 30
8. The routing apparatus of any of claims 1 to 7, further comprising means for determining from the packet, if it has previously triggered a sending, to the source node, of an indication of congestion, the output means being operable according to whether such an indication had been sent previously. 35
9. A source apparatus for use in a packet network comprising a plurality of nodes, the apparatus comprising: 40
- output means for sending a packet via an intermediate one of the nodes in the network to a receiving one of the nodes;
- input means for receiving from the intermediate one of the nodes an indication of a degree of congestion at that intermediate one, sent using an OSI network layer protocol, in response to the packet; and
- a controller for controlling a flow of further packets to the receiving node on the basis of the indication. 45
10. A method of processing packets in a network comprising a receiver node, a source node and at least one intermediate routing node, the method comprising the steps of: at one of the intermediate routing nodes, receiving a packet passed across the 50
- network;
- determining a degree of congestion at the routing node; and
- sending an indication of this degree of congestion to the source node, using an OSI network layer protocol. 55
11. The method of claim 10, further comprising the step of, at the source node:
- receiving from the intermediate routing node the indication of the degree of congestion; and
- controlling a flow of further packets to the receiving node on the basis of the indication. 60
12. A method of processing packets in a network comprising a receiver node, a source node and at least one intermediate routing node, the method comprising the steps of, at the source node:
- sending a packet to the receiver node;
- receiving from the intermediate routing node an indication of a degree of congestion at that intermediate routing node, sent using an OSI network layer protocol, in response to the packet; and
- controlling a flow of further packets to the receiving node on the basis of the indication. 65
13. The method of claim 12, further comprising the step of receiving from the receiver node a flow control message, the step of controlling the flow of further packets being made also on the basis of this flow control message. 70
14. Packet routing apparatus for routing packets in a network comprising a plurality of nodes, the routing apparatus comprising: 75
- an input for receiving a packet passed across the network;
- a congestion monitor for determining congestion in the routing apparatus;
- a packet reader for determining from the packet, if it has previously triggered a sending, to the source node, of an indication of congestion at another of the intermediate routing nodes; and
- an output for sending an indication of congestion to the source node according to whether such an indication had been sent previously. 80
15. A method of processing packets in a network comprising a receiver node, a source node and at least one intermediate routing node, the method comprising the steps of, at one of the intermediate routing nodes:

- receiving a packet passed across the network; determining congestion in the routing apparatus;
determining from the packet, if it has previously triggered a sending, to the source node, of an indication of congestion at another of the intermediate routing nodes; and
sending an indication of the congestion to the source node according to whether such an indication had been sent previously.
16. The method of claim 15, the packet being an Internet Protocol packet, and the step of determining from the packet, if it has previously triggered a sending of an indication comprising checking the Congestion Experienced bit in the packet header.
17. The method of claim 15, the step of sending an indication not being carried out if the indication had been sent previously, unless the routing apparatus discards the packet.
18. A method of using an application to transmit data across a network, the application causing the network to use a packet routing apparatus as set out in claim 1, to transmit the data.
19. Software stored on a computer readable medium for carrying out the method of any of claims 10 or 12.
20. A receiver apparatus for use in an Internet Protocol network comprising a plurality of nodes, the apparatus comprising:

input means for receiving a packet sent across the network using the Internet Protocol;
packet reading means for determining if the packet has been marked according to the Internet Protocol by any of the nodes through which it passed, to indicate congestion at that node;
a packet flow control parameter generator responsive to the packet reading means, for determining a packet flow control parameter; and
output means for sending a message to the source, to control the flow of packets from the source, according to the packet flow control parameter.
21. A source node apparatus for use in an Internet Protocol network comprising a plurality of nodes, the apparatus comprising:

a packet sending means for sending a packet using across the Internet Protocol network to one of the nodes acting as a receiving node;
a receiver for receiving from the receiving node a packet flow control parameter sent across the Internet Protocol network; and
- 5 22. The source node apparatus of claim 21 the receiver being capable of receiving congestion notification messages from other nodes in the path of the packet between the source node and the receiving node, the flow controller being operable to control the rate of flow on the basis of the congestion notification messages received from the other nodes.
- 15 23. A method of processing packets in an Internet Protocol network comprising a receiver node, a source node and intermediate nodes, the method comprising the steps of: at the receiver node,

receiving a packet sent across the network using the Internet Protocol;
determining if the packet has been marked by any of the nodes through which it passed, to indicate congestion at that node;
determining a packet flow control parameter on the basis of the determination of a marking; and
sending a message to the source node across the Internet Protocol network, to control the flow of packets from the source node, according to the packet flow control parameter.
- 30 24. The method of claim 23 further comprising the steps of: at the source node,

receiving the message; and
controlling a rate of flow of sending further packets to the receiving node on the basis of the message.
- 35 40 45 50 55

FIG. 1 PRIOR ART

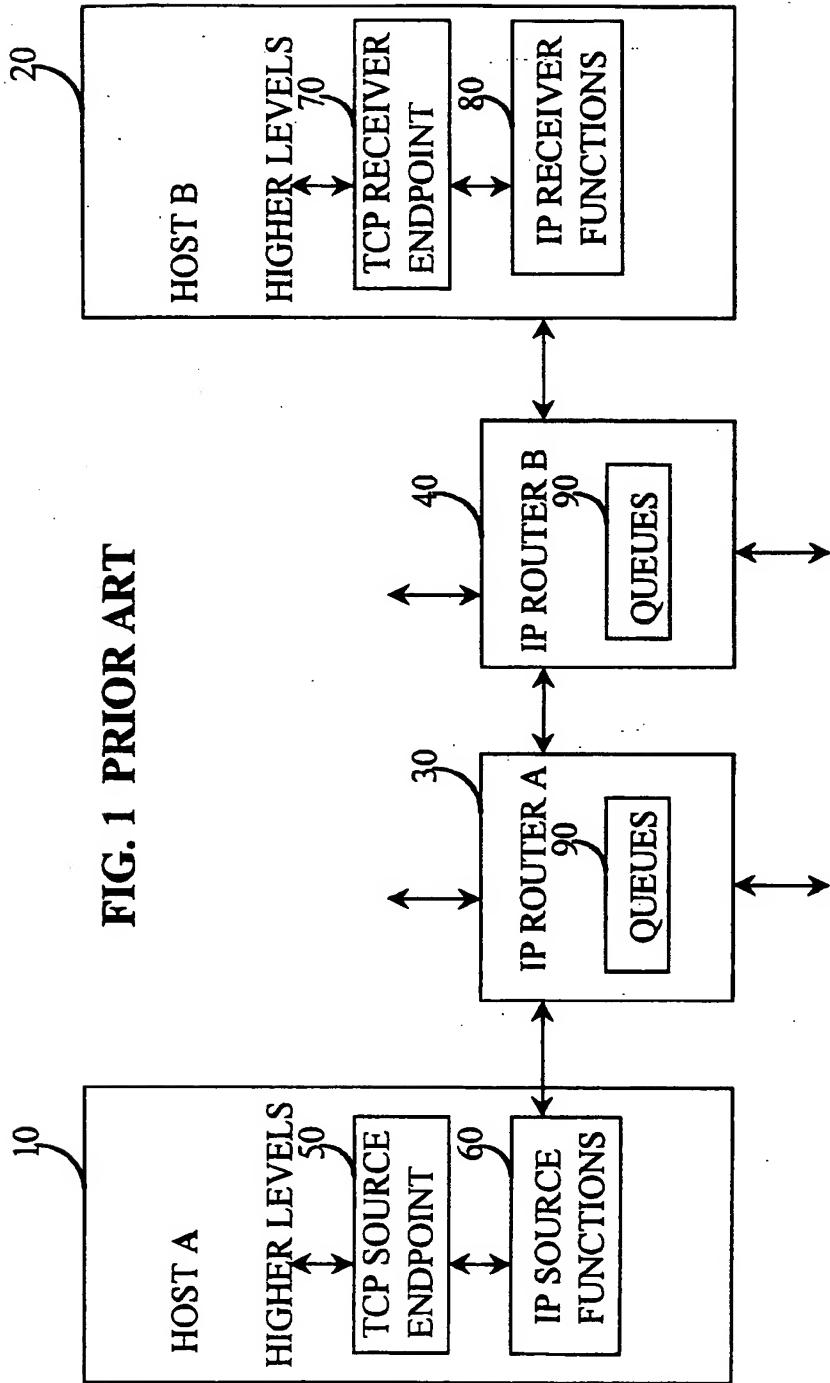


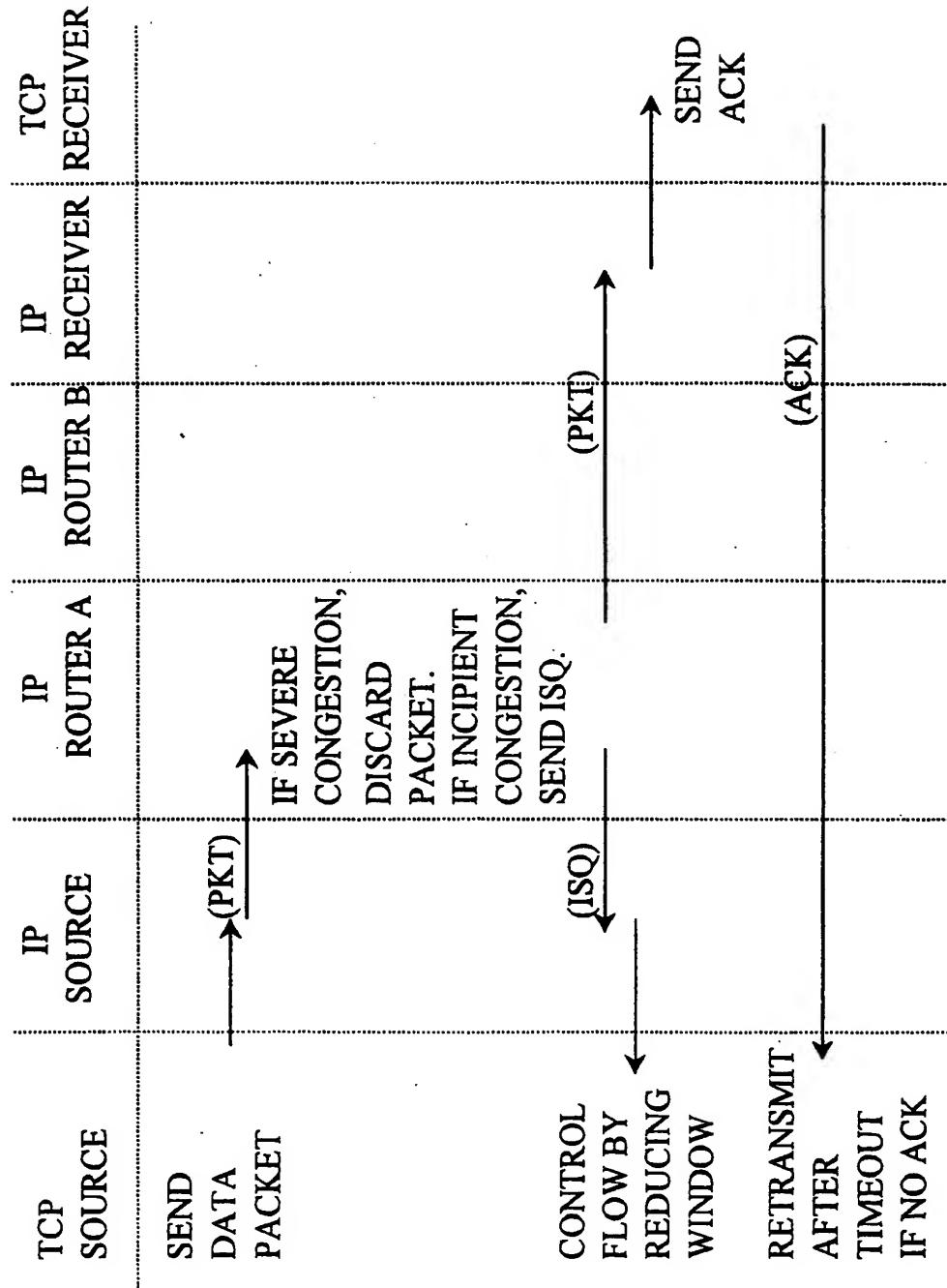
FIG. 2 PRIOR ART

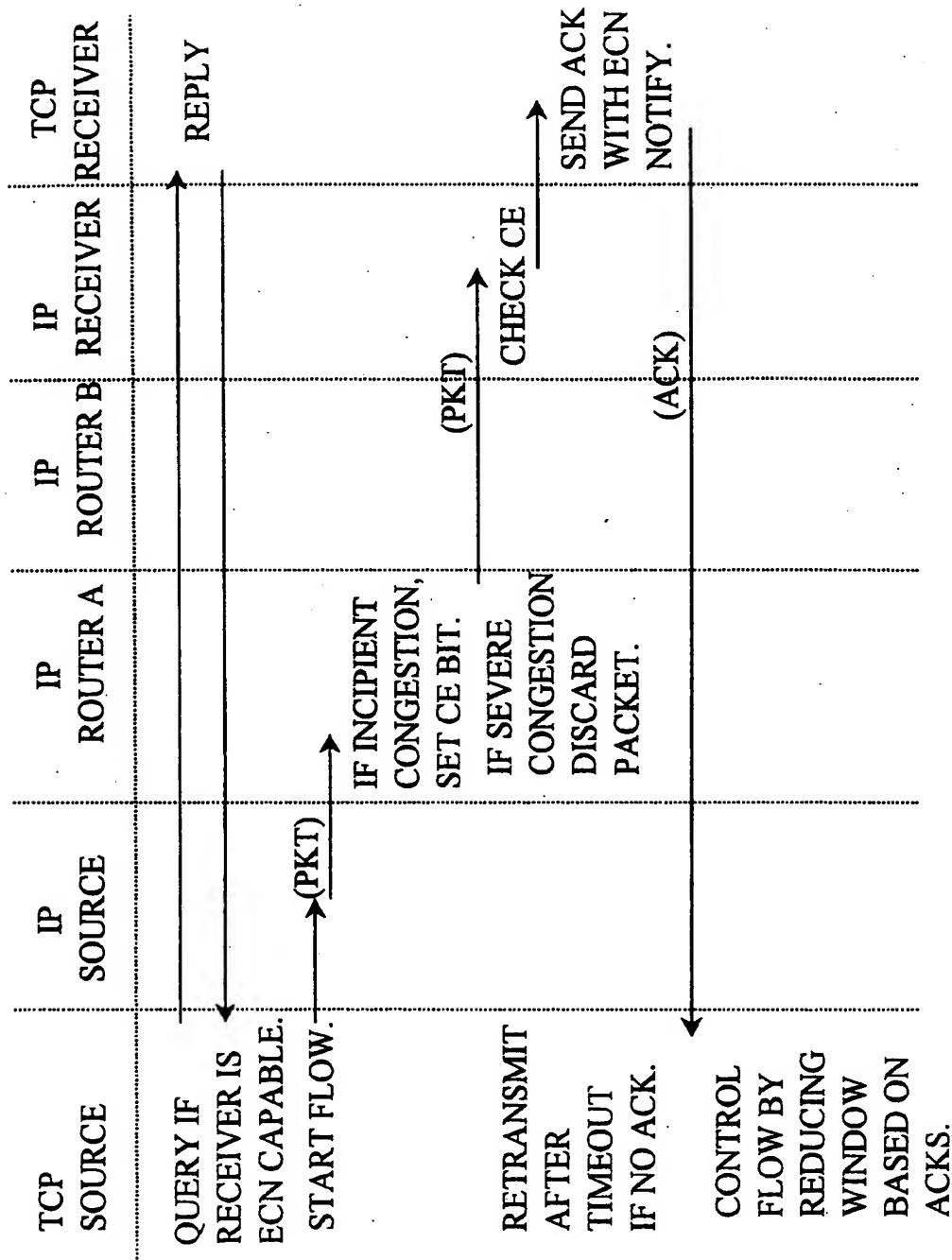
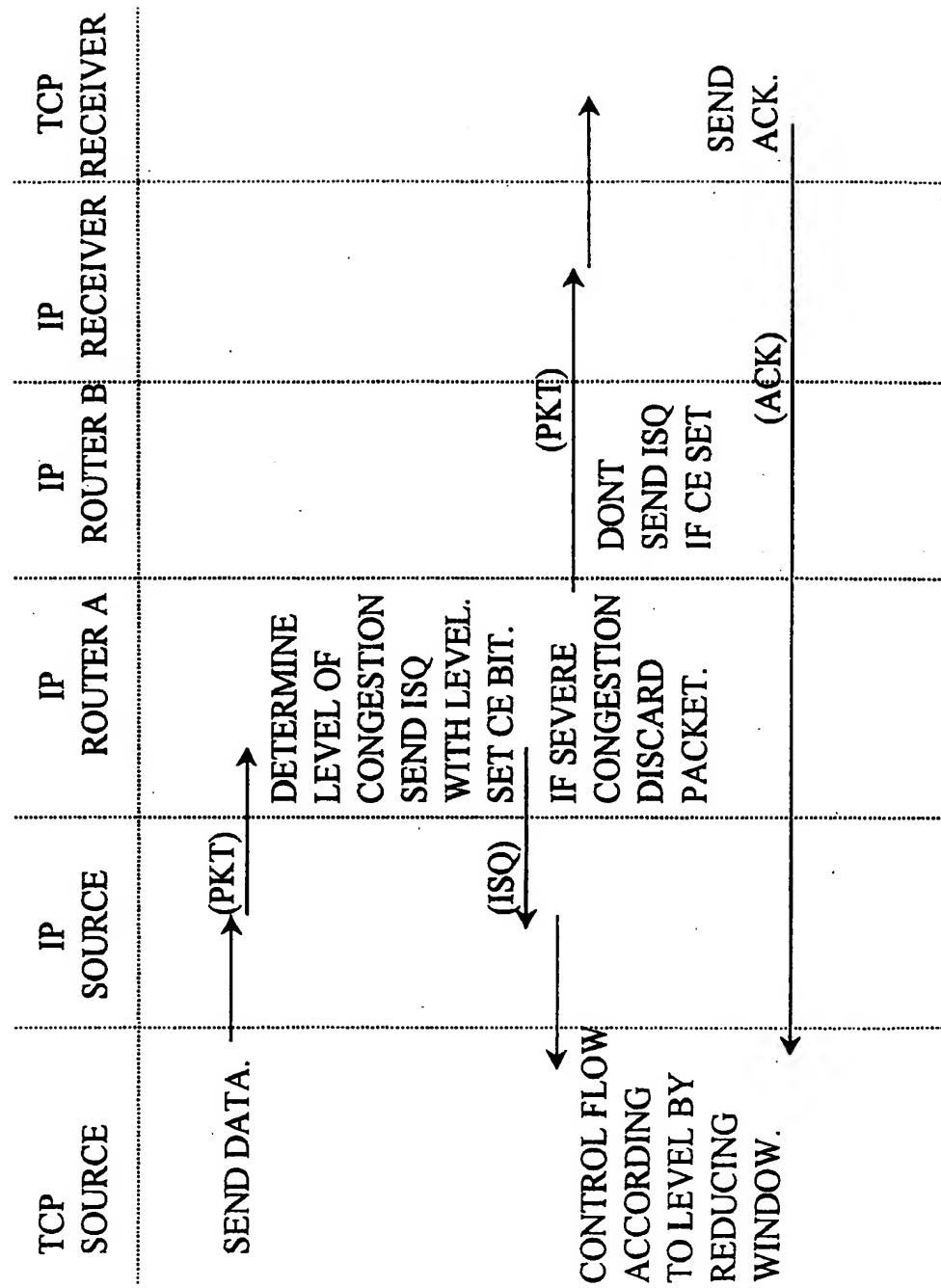
FIG. 3 PRIOR ART

FIG. 4

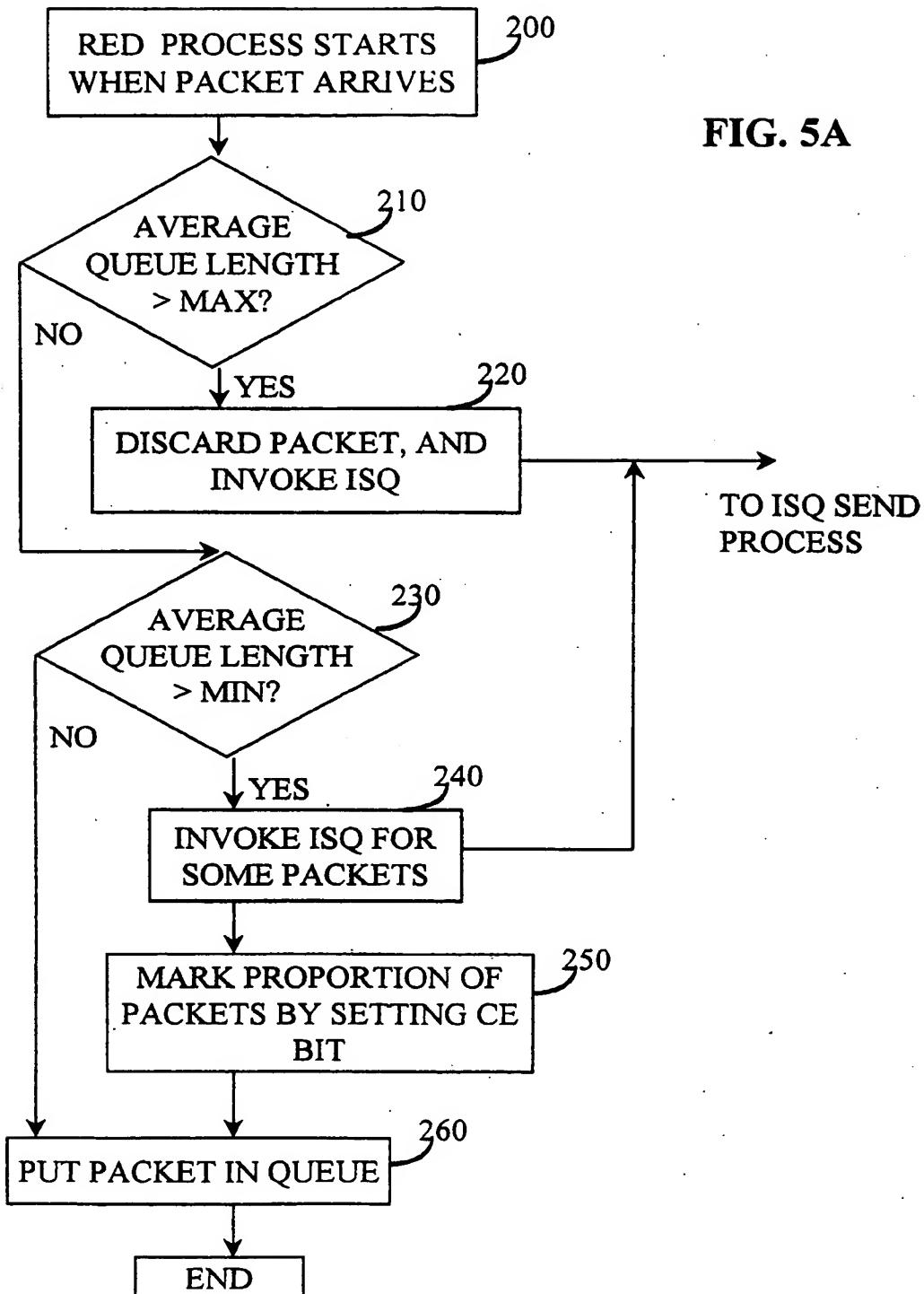
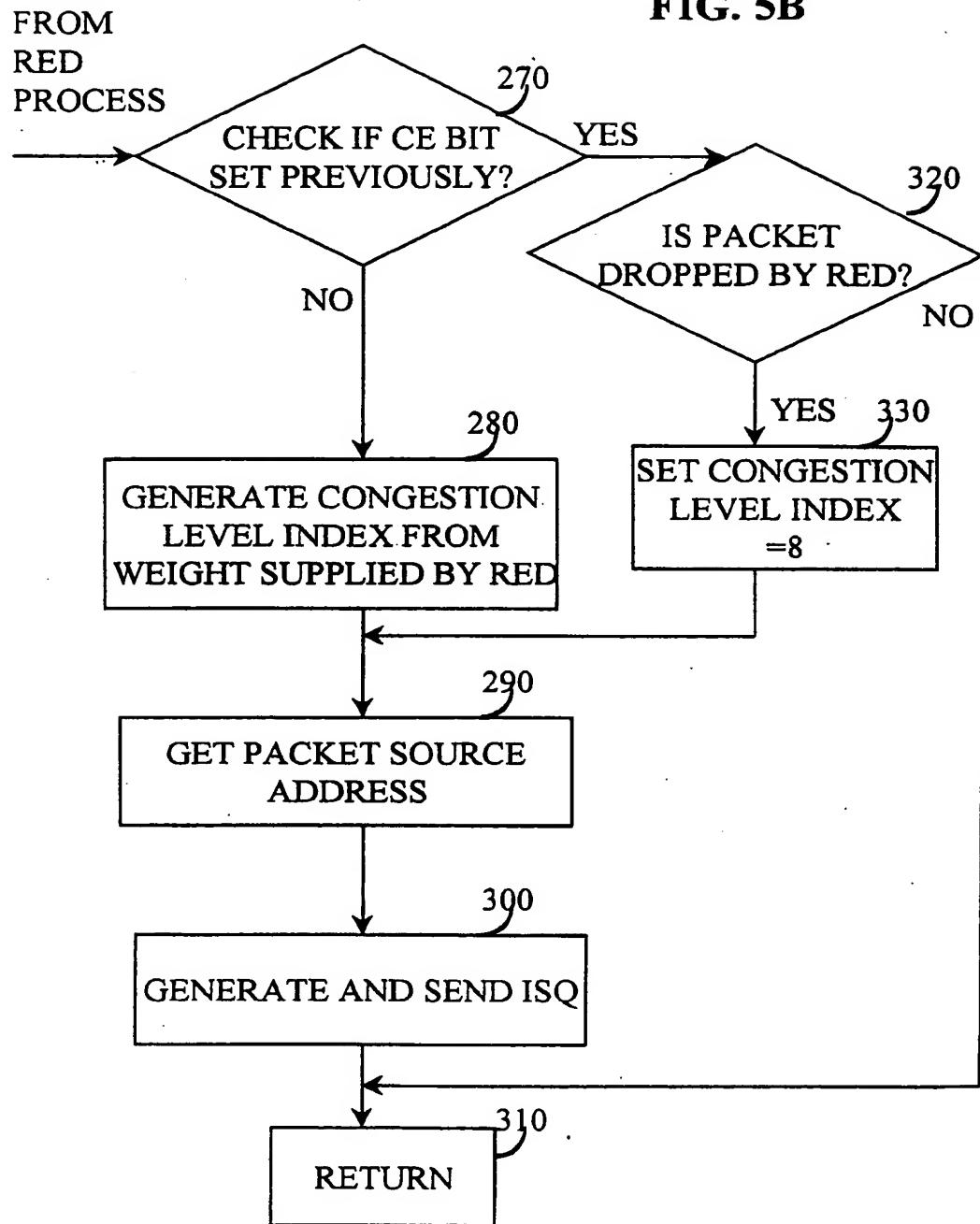


FIG. 5B

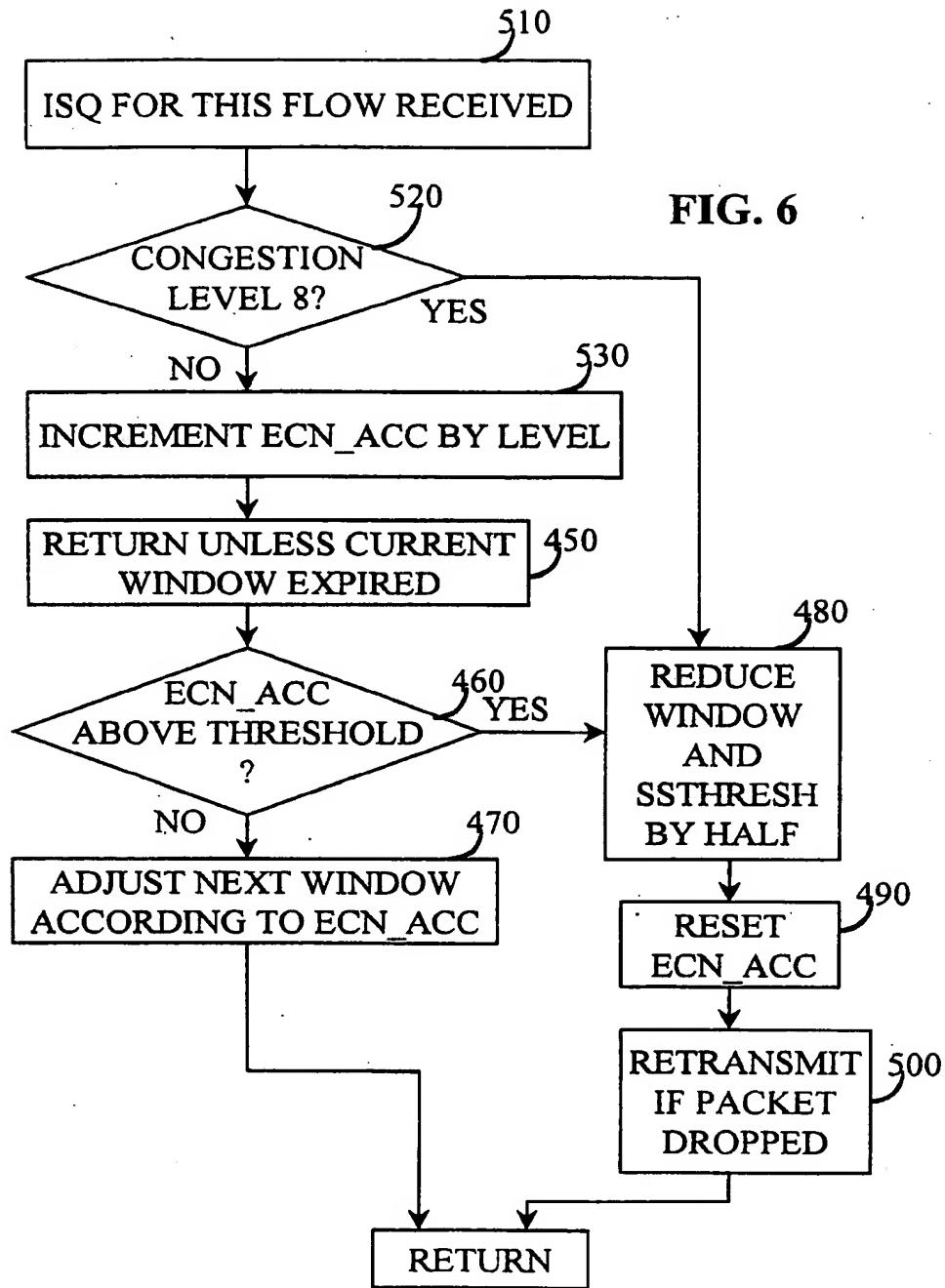


FIG. 7

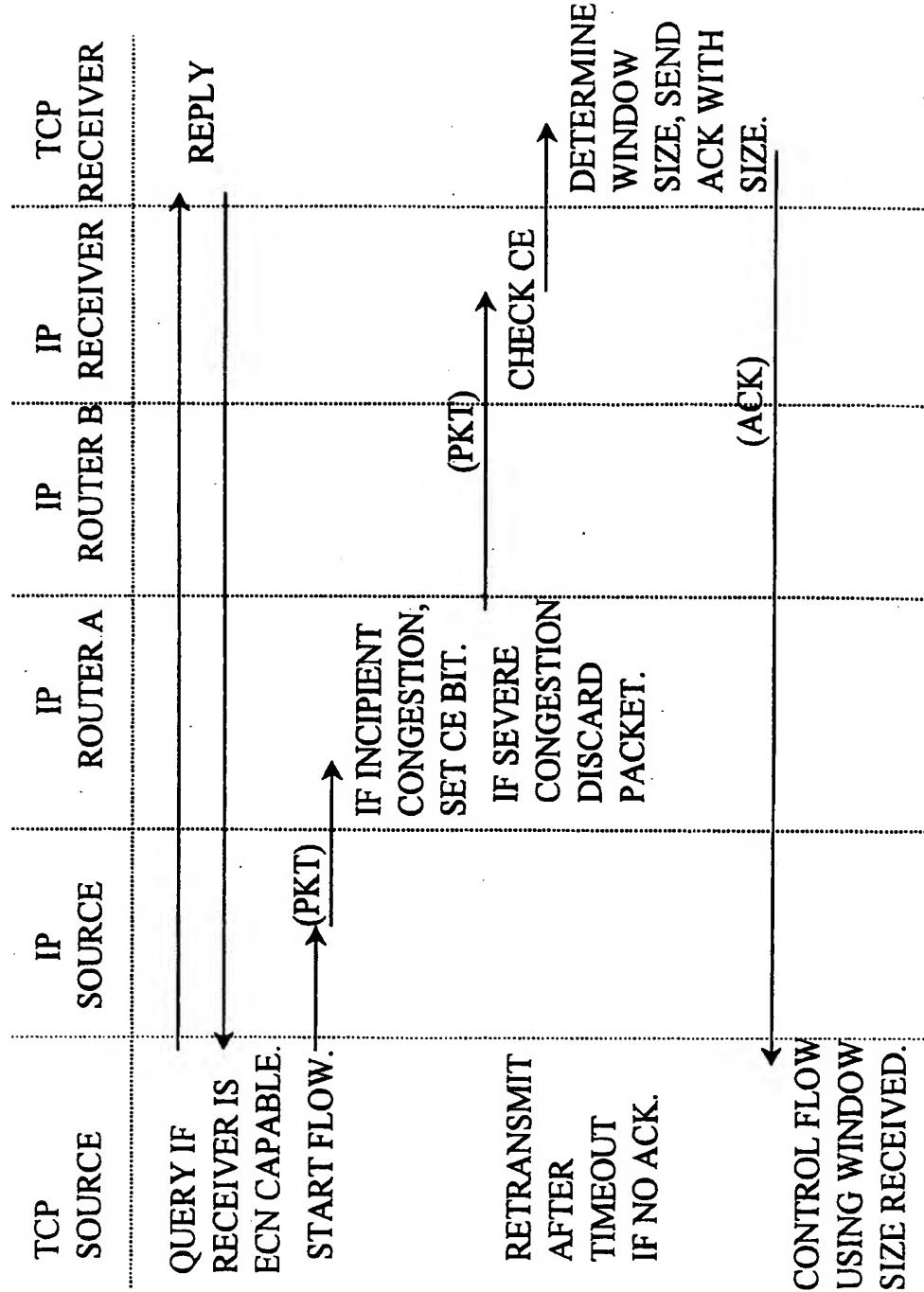
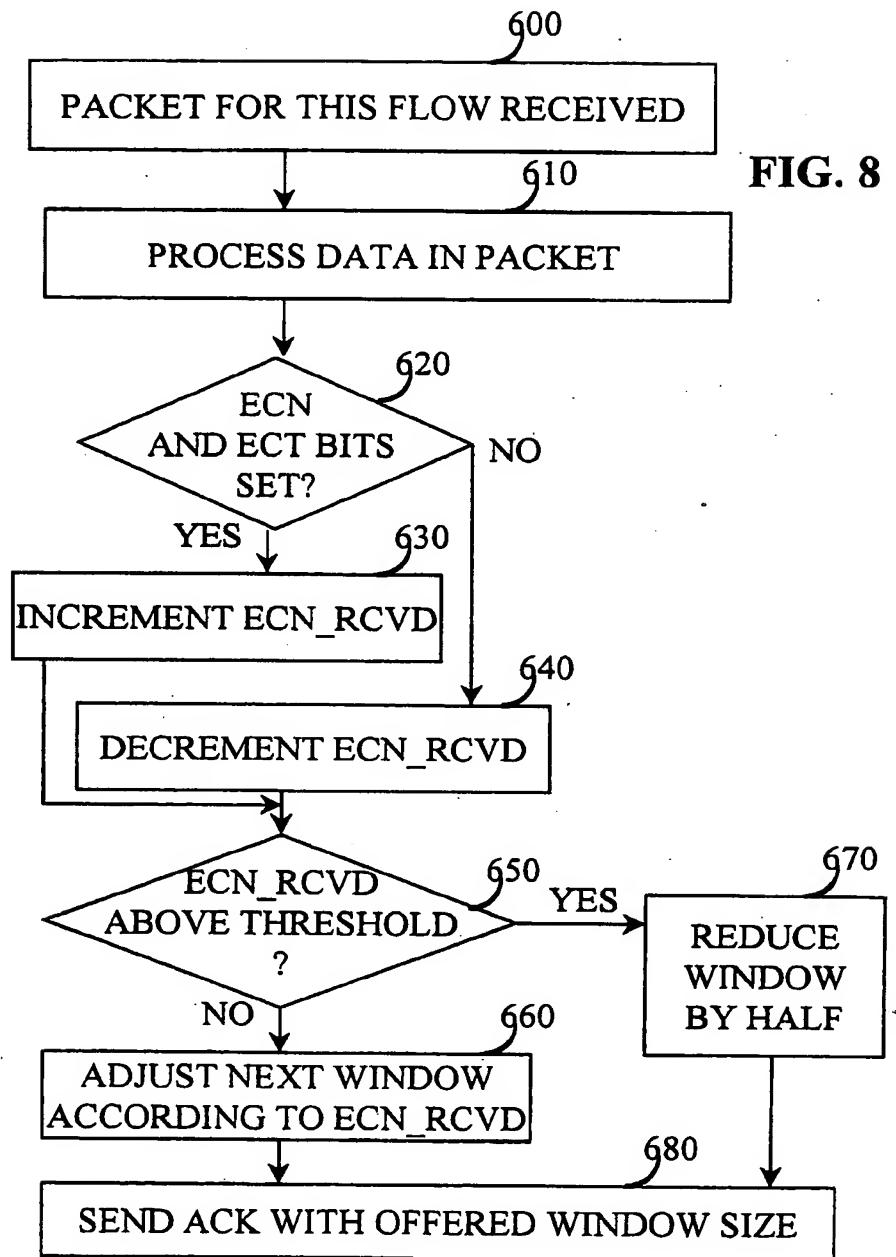


FIG. 8



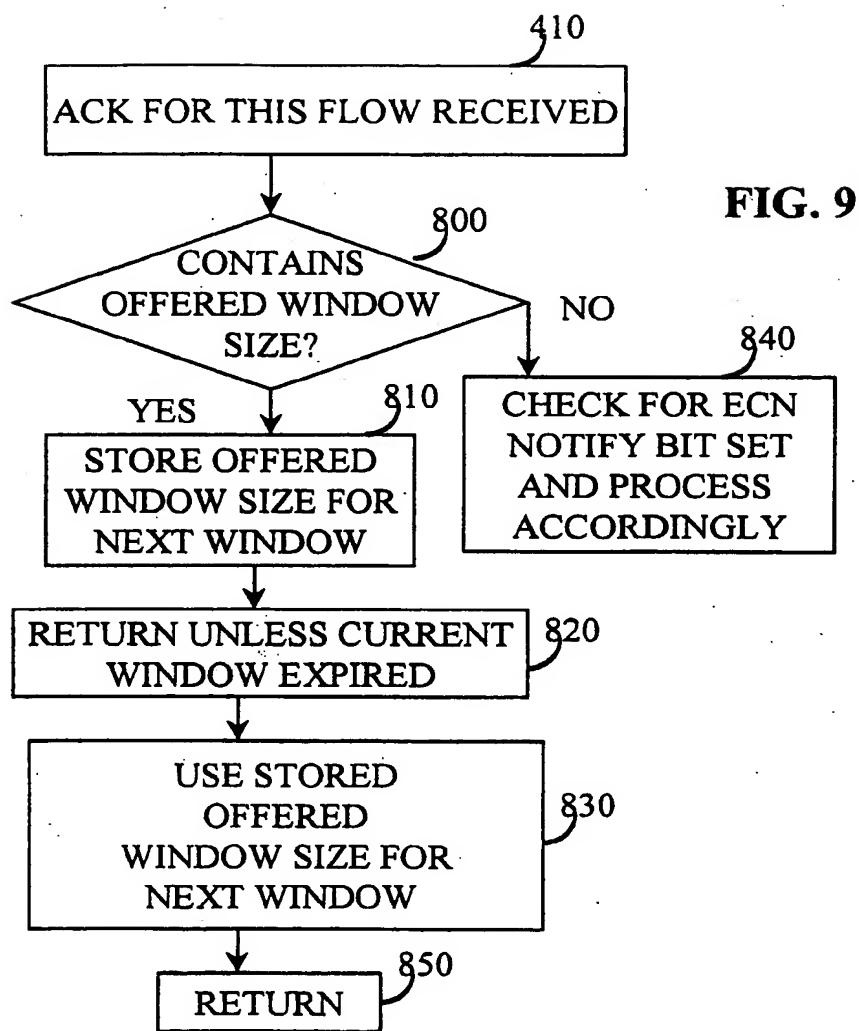


FIG. 10

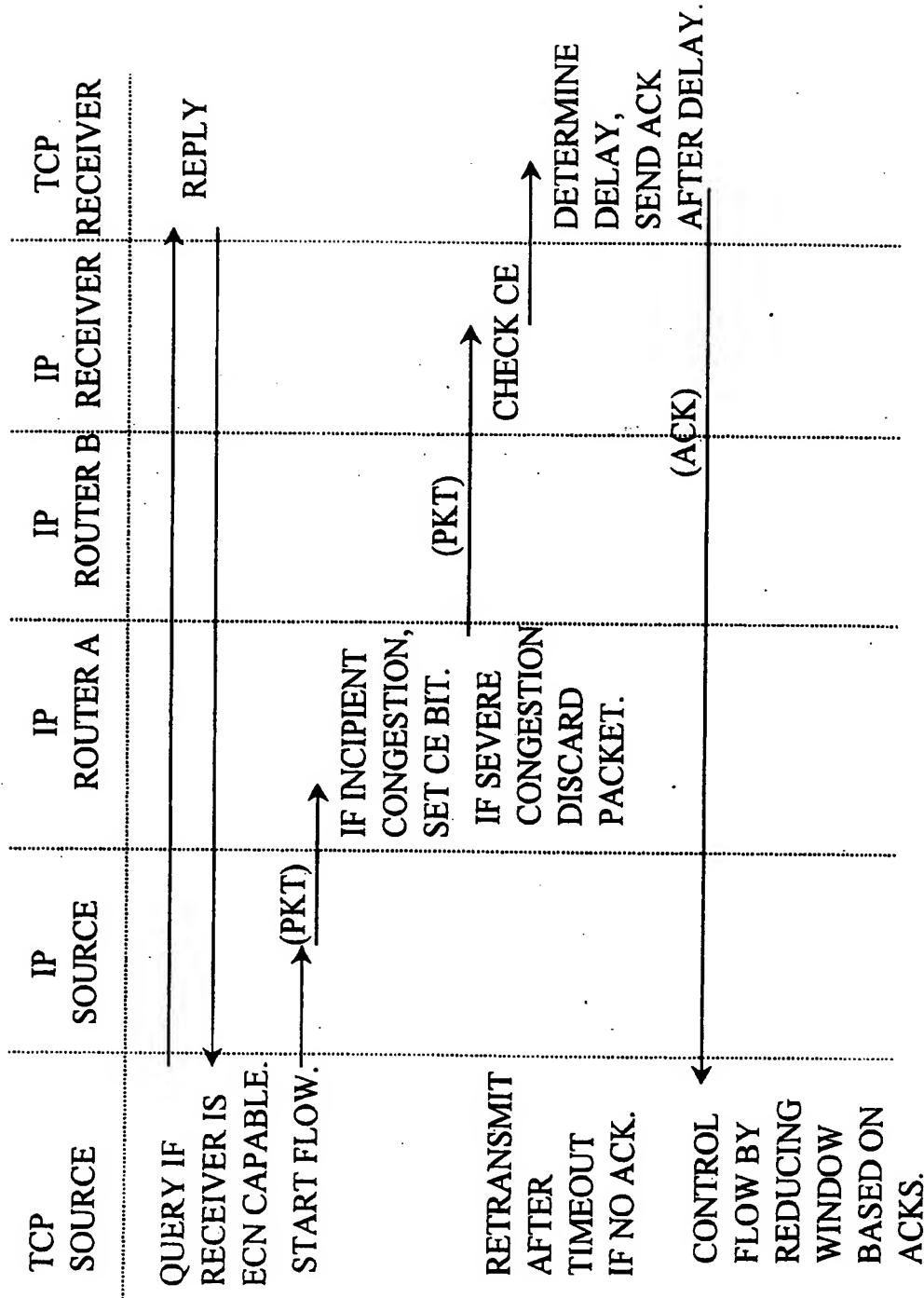
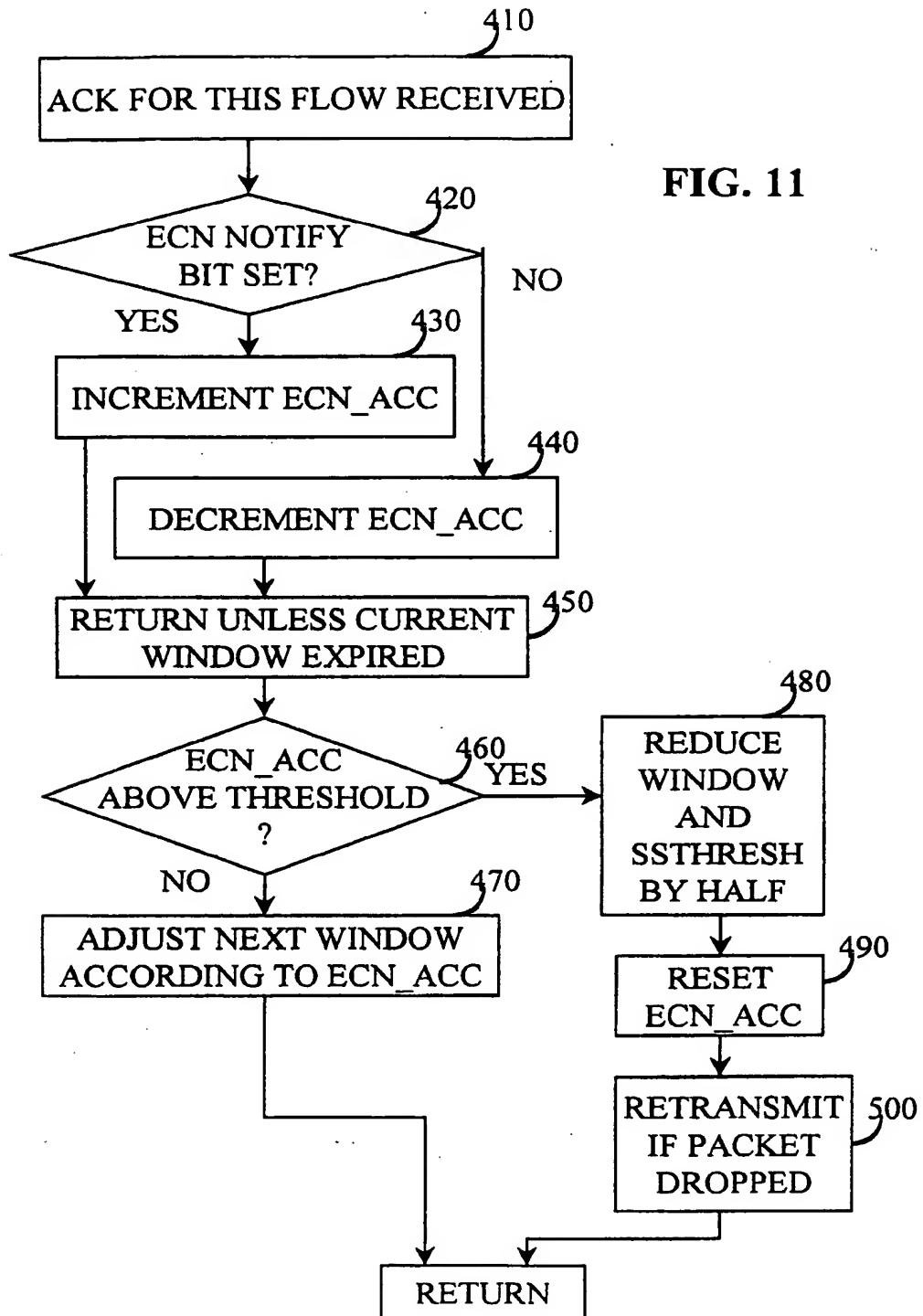


FIG. 11





European Patent
Office

EUROPEAN SEARCH REPORT

Application Number
EP 99 30 3353

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.)
X	US 5 313 454 A (DALEY PATRICK D ET AL) 17 May 1994 (1994-05-17) * abstract * * figure 1 * * column 1, line 19 - line 32 * * column 5, line 17 - line 41 *	1,4,6,9, 10,12, 19,20	H04L12/56 H04L12/18
TECHNICAL FIELDS SEARCHED (Int.Cl.)			
H04L			
The present search report has been drawn up for all claims			
Place of search	Date of completion of the search	Examiner	
THE HAGUE	14 September 1999	Adkhis, F	
CATEGORY OF CITED DOCUMENTS			
X : particularly relevant if taken alone Y : particularly relevant if combined with another document of the same category A : technological background O : non-written disclosure P : intermediate document			
T : theory or principle underlying the invention E : earlier patent document, but published on, or after the filing date D : document cited in the application L : document cited for other reasons & : member of the same patent family, corresponding document			

**ANNEX TO THE EUROPEAN SEARCH REPORT
ON EUROPEAN PATENT APPLICATION NO.**

EP 99 30 3353

This annex lists the patent family members relating to the patent documents cited in the above-mentioned European search report.
The members are as contained in the European Patent Office EDP file on
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

14-09-1999

Patent document cited in search report	Publication date	Patent family member(s)		Publication date
US 5313454 A	17-05-1994	AU	3797093 A	08-11-1993
		EP	0634068 A	18-01-1995
		JP	7508378 T	14-09-1995
		WO	9320637 A	14-10-1993

THIS PAGE BLANK (USP70)